Celligner multi-dataset alignment documentation

Overview

In the Celligner manuscript (see here), we demonstrated a method for aligning two RNA-Seq datasets: one from patient tumor samples and one from cancer cell lines. For the Celligner app, we've now expanded this to include several additional datasets (one of metastatic tumor samples, and two PDX datasets). Aligning additional datasets with Celligner requires making some additional steps in the algorithm, and we're still exploring the optimal ways to do this. We describe the datasets used, and the modifications we've made to the Celligner method, in more detail below.

Datasets Used

- **Cancer cell lines (DepMap)**
  - We've updated the cell line dataset to the 21Q1 DMC data. The cell line data used as input can be found at the DepMap portal.
  - The file is DepMap DMC 21Q1 CCLE_expression_full.csv
- **Treehouse tumors (TCGA+)**
  - This compilation of TCGA/Treehouse tumor RNA-Seq profiles is the same as used in the Celligner paper.
  - We used Treehouse Tumor Compendium v10 Public PolyA, available from UCSC Xena here: https://xenabrowser.net/datapages/?dataset=TumorCompendium_v10_PolyA_hugo_log2tpm_58581genes_2019-07-25.tsv&host=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443

New datasets

- **Metastatic tumors (Met500)**
  - Data from 868 metastatic tumor samples are from Robinson et al., Nature 2017: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5995337/.
  - We downloaded the RNA-Seq data from UCSC Xena https://xenabrowser.net/datapages/?cohort=MET500%20(expression%20centric)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443
  - We then converted the data from FPKM to TPM
- **Pediatric PDX**
  - Data for 244 pediatric PDX models are from Rokita et al., Cell Reports (2019): https://www.sciencedirect.com/science/article/pii/S2211124719312689#mmc2
  - We then converted the data from FPKM to TPM
- **Novartis PDX Encyclopedia**
  - Data from 446 PDX samples are taken from the Novartis PDXE dataset (Gao et al., 2015, https://www.nature.com/articles/nm.3954).

    - Specifically, we used a version of the data processed using PISCES (https://www.biorxiv.org/content/10.1101/2020.12.01.390575v1), which in turn uses Salmon. This dataset, created by Joshua Korn, is available on Figshare https://figshare.com/articles/dataset/pdxe_sample_annotations_txt/13331072
    - These data were then converted from gene-level counts to TPM for Celligner analysis

Lightweight metadata cleanup:

- The disease annotations for the “lineage” and “subtype” columns differed slightly across the data sources. We performed lightweight cleanup on these columns to harmonize the metadata.
    - The mapping we used is available here: https://docs.google.com/spreadsheets/d/1Vc1XQRDU7kzD0zaDD0BLX7XAkTw71zYxESIOpmWgey0/edit?usp=sharing
- We also harmonized the annotations for the “Primary/Metastatic” column so that the only unique values are “Primary”, “Metastatic”, “Recurrent Tumor”, or NA. This involved obvious changes like mapping “Primary Tumor” to “Primary”. Note that the only datasets we included primary vs metastatic information for were TCGA, DepMap, and Met500.

Multi-dataset alignment procedure

To perform alignment of the above RNA-Seq datasets, we first run the Celligner method as described in the paper to align the Treehouse/TCGA and DepMap datasets. We then sequentially align the Met500, Novartis PDX, and Pediatric PDX data using the following procedure:

- We first regress out the contrastive PCs calculated by comparing the Treehouse/TCGA + DepMap datasets
- We perform mutual nearest neighbors (MNN) alignment using the differentially expressed genes calculated from the Treehouse/TCGA + DepMap data, as in the Celligner paper.
- Each sequentially added via MNN alignment in the order specified above
- The ‘k’ (number of nearest neighbors) parameters in the MNN alignment were set at 20, 10, and 10 for the Met500, Novartis PDX, and Pediatric PDX datasets respectively. In all cases we used k=50 for the larger reference dataset. These parameter values were chosen ‘by hand’ to roughly scale by the relative size of each dataset. In general we’ve found that k-values between 5 and 50, depending on the dataset size, produce the best results.